

---

# A Variational Bayesian State-Space Approach to Online Passive-Aggressive Regression

---

Arnold Salas\*, Stephen J. Roberts and Michael A. Osborne

Department of Engineering Science and Oxford-Man Institute

University of Oxford

arnold.salas@eng.ox.ac.uk, {sjrob, mosb}@robots.ox.ac.uk

## Abstract

Online Passive-Aggressive (PA) learning is a class of online margin-based algorithms suitable for a wide range of real-time prediction tasks, including classification and regression. PA algorithms are formulated in terms of deterministic point-estimation problems governed by a set of user-defined hyperparameters: the approach fails to capture model/prediction uncertainty and makes their performance highly sensitive to hyperparameter configurations. In this paper, we introduce a novel PA learning framework for regression that overcomes the above limitations. We contribute a Bayesian state-space interpretation of PA regression, along with a novel online variational inference scheme, that not only produces probabilistic predictions, but also offers the benefit of automatic hyperparameter tuning. Experiments with various real-world data sets show that our approach performs significantly better than a more standard, linear Gaussian state-space model.

## 1 Introduction

Online learning is the most common approach of learning from non-stationary and/or large sequential data sets. In online learning, model parameters are learned in a sequential manner, thus achieving temporal adaptation and learning efficiency in time-aware applications. Among the popular algorithms, online Passive-Aggressive (PA) learning [1] provides a generic family of online margin-based algorithms for various time-aware applications, including classification and regression. However, despite their merits, PA algorithms make point rather than probabilistic predictions, and depend on a set of hyperparameters that are assumed to be user-defined and constant over time. This assumption is impractical for at least two reasons. First, it has been recently argued that the performance of many machine learning algorithms is highly sensitive to hyperparameter settings [2], and PA learning is unlikely to be an exception because its performance is measured in terms of cumulative loss. Second, in non-stationary environments, optimal hyperparameter choices may quickly become sub-optimal, due to the evolving nature of the underlying population distributions.

To address these drawbacks, we propose a new online PA method based on a Bayesian treatment of the existing PA framework. We concentrate here on PA learning for regression. Our algorithm incorporates a novel, online, variational inference scheme. Furthermore, it explicitly takes into account uncertainty in our predictions and is endowed with a self-tuning hyperparameter mechanism.

The main contributions of the paper are twofold. Firstly, this paper is, to the best of our knowledge, the first to approach online PA regression from a Bayesian state-space perspective. We will indeed show that the state-space representation of PA regression results in a Bayesian linear Gaussian state-space model (LGSSM). Secondly, we establish a clear connection between our online variational

---

\*Corresponding author. This work was funded through AFR-PhD grant agreement 8837255 from the National Research Fund of Luxembourg, and by the Economic and Social Research Council (ESRC) and the Oxford-Man Institute.

inference procedure and Streaming Variational Bayes [3], thus making the first application of the latter to the Bayesian LGSSM setting.

## 2 Bayesian State-Space Approach to Passive-Aggressive Regression

In this section, we provide a Bayesian treatment of online PA regression within a state-space framework. We show that the state-space model (SSM) corresponding to PA regression is, conditionally upon the mean and variance of the measurement noise, a special case of the Bayesian LGSSM, and that it justifies the PA regression algorithm from a *maximum a posteriori* (MAP) standpoint.

### 2.1 Online Passive-Aggressive Regression

Consider a data stream consisting of examples  $\{(\mathbf{x}_t, y_t)\}_{t \geq 1}$ , where  $\mathbf{x} \in \mathbb{R}^I$  is an  $I$ -dimensional input vector and  $y \in \mathbb{R}$  is the associated output. Online PA regression [1] is based on the linear prediction model of the form  $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}$ , where  $\mathbf{w} \in \mathbb{R}^I$  is the incrementally learned weight vector. The PA regression algorithm initialises the weight vector to the zero vector ( $\hat{\mathbf{w}}_1 = \mathbf{0}_{I \times 1}$ ) and, after observing the  $t^{\text{th}}$  example, the new weight  $\hat{\mathbf{w}}_t$  is obtained as the solution to<sup>1</sup>

$$\min_{\mathbf{w}_t} \left\{ \frac{1}{2} \|\mathbf{w}_t - \hat{\mathbf{w}}_{t-1}\|_2^2 + C \ell(y_t, \mathbf{x}_t^\top \mathbf{w}_t; \epsilon) \right\}, \quad (1)$$

where  $\ell(y, \hat{y}; \epsilon) = |y - \hat{y}|_\epsilon \equiv \max(|y - \hat{y}| - \epsilon, 0)$  is the  $\epsilon$ -insensitive loss function ( $\epsilon$ -ILF) and  $C > 0$  is a user-specified parameter. The intuitive goal of PA regression is to minimally change the existing weight estimate while predicting the  $t^{\text{th}}$  example as accurately as possible. The parameter  $C$  serves to balance these two competing objectives. Larger values of  $C$  imply a more aggressive update step, whence the name of *aggressiveness parameter* [1].

### 2.2 Bayesian Linear Gaussian State-Space Models

LGSSMs<sup>2</sup> are fundamental in time-series analysis [4, 5]. In these models, each output  $y_t$  is generated from an underlying dynamical system on the hidden variable  $\mathbf{h}_t$  according to:

$$y_t = \mathbf{b}^\top \mathbf{h}_t + \eta_t, \quad \eta_t \sim \mathcal{N}(\eta_t | 0, \sigma^2), \quad \mathbf{h}_t = \mathbf{A} \mathbf{h}_{t-1} + \boldsymbol{\eta}_t^{\mathbf{h}}, \quad \boldsymbol{\eta}_t^{\mathbf{h}} \sim \mathcal{N}(\boldsymbol{\eta}_t^{\mathbf{h}} | \mathbf{0}_{H \times 1}, \boldsymbol{\Sigma}), \quad (2)$$

where  $H \equiv \dim(\mathbf{h}_t)$ . The initial latent variable also has a Gaussian distribution which we write as  $p(\mathbf{h}_1) = \mathcal{N}(\mathbf{h}_1 | \boldsymbol{\mu}_\pi, \boldsymbol{\Sigma}_\pi)$ . The model parameters are therefore  $\boldsymbol{\theta} \equiv (\mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma}, \sigma^2, \boldsymbol{\mu}_\pi, \boldsymbol{\Sigma}_\pi)$ . In the Bayesian treatment of the LGSSM, instead of considering  $\boldsymbol{\theta}$  as fixed, we define a prior distribution  $p(\boldsymbol{\theta} | \boldsymbol{\omega})$ , where  $\boldsymbol{\omega}$  is a vector of hyperparameters.

### 2.3 Bayesian State-Space Representation of Passive-Aggressive Regression

Let  $\mathbf{I}_I$  be the identity matrix of order  $I$ . The state-space representation of PA regression is given by

$$y_t = \mathbf{x}_t^\top \mathbf{w}_t + \eta_t, \quad \eta_t \sim p(\eta_t | \epsilon), \quad \mathbf{w}_t = \mathbf{w}_{t-1} + \boldsymbol{\eta}_t^{\mathbf{w}}, \quad \boldsymbol{\eta}_t^{\mathbf{w}} \sim \mathcal{N}(\boldsymbol{\eta}_t^{\mathbf{w}} | \mathbf{0}_{I \times 1}, \alpha^{-1} \mathbf{I}_I), \quad (3)$$

with the convention that  $\mathbf{w}_0 = \mathbf{0}_{I \times 1}$ , and where

$$p(\eta_t | \epsilon) = \frac{1}{2(1 + \epsilon)} e^{-|\eta_t|_\epsilon} \quad (4)$$

is the measurement-noise density dictated by the  $\epsilon$ -ILF [6]. In this case, the weight posterior satisfies<sup>3</sup>

$$p(\mathbf{w}_t | y_t, \mathbf{w}_{t-1}) \propto p(y_t | \mathbf{w}_t) p(\mathbf{w}_t | \mathbf{w}_{t-1}) = \frac{1}{2(1 + \epsilon)} e^{-\ell(y_t, \mathbf{x}_t^\top \mathbf{w}_t; \epsilon)} \mathcal{N}(\mathbf{w}_t | \mathbf{w}_{t-1}, \alpha^{-1} \mathbf{I}_I). \quad (5)$$

Setting  $(\alpha, \mathbf{w}_{t-1}) = (C^{-1}, \hat{\mathbf{w}}_{t-1})$  in the above equation, taking the negative logarithm thereof and ignoring any resulting additive constant yields the PA objective from (1). We thus obtain a MAP justification for the PA regression algorithm.

<sup>1</sup>We restrict our attention to the PA-I variant of PA regression.

<sup>2</sup>These are also called Kalman Filters/Smoothers and Linear Dynamical Systems.

<sup>3</sup>For brevity, we omit  $\mathbf{x}_t$  from the conditioning statements, and shall do so in the remainder of the paper.

Observe that Eqs. (3)-(4) give a model that is intractable, due to the Laplacian-like noise distribution. Having said that, [7] proved that this distribution can be expressed as a continuous mixture of Gaussians (CMoG). Specifically<sup>4</sup>,

$$p(\eta_t|\epsilon) = \int \int \mathcal{N}(\eta_t|\mu, \beta^{-1}) p(\mu|\epsilon) p(\beta) d\mu d\beta, \quad (6)$$

with

$$p(\beta) = \mathcal{IG}(\beta|1, 1/2) = \frac{1}{2} \beta^{-2} e^{-\frac{1}{2\beta}} \quad (7)$$

$$p(\mu|\epsilon) = \overline{\mathcal{U}}(\mu|-\epsilon, \epsilon) \equiv \frac{1}{2(1+\epsilon)} [\mathbb{1}_{[-\epsilon, \epsilon]}(\mu) + \delta(\mu + \epsilon) + \delta(\mu - \epsilon)], \quad (8)$$

where  $\mathcal{IG}$  stands for ‘inverse Gamma’,  $\mathbb{1}_S(\cdot)$  for the indicator function of the set  $S$ , and  $\delta(\cdot)$  for the Dirac delta function. The above CMoG formulation implies that, conditionally upon  $\beta$  and  $\mu$ , the SSM described by Eqs. (3)-(4) is a special case of the Bayesian LGSSM from (2). To retain this formalism, we will, in the first instance, hold  $\beta$  and  $\mu$  ‘fixed’. In the second instance, we will approximately marginalise  $\beta$  and  $\mu$  by means of an innovative, truly sequential, Variational Bayes (VB) routine.

Going forward, we shall refer to the ensuing model as *Bayesian Passive-Aggressive State-Space Model*, or BYPASS for short. BYPASS additionally takes the prior over its parameter vector  $\theta = (\alpha, \beta, \mu)$  to factorise as

$$p(\theta|\omega) = \mathcal{G}(\alpha|a, b) \mathcal{IG}(\beta|1, 1/2) \overline{\mathcal{U}}(\mu|-\epsilon, \epsilon), \quad \omega = (a, b, \epsilon). \quad (9)$$

Note that we have assigned the standard conjugate prior to the weight precision  $\alpha$ . We do not define any prior for  $\omega$ <sup>5</sup>. Probabilistically, the BYPASS model is defined by<sup>6</sup>

$$p(y_{1:t}, \mathbf{w}_{1:t}, \theta|\omega) = p(y_{1:t}, \mathbf{w}_{1:t}|\theta) p(\theta|\omega) = \left[ \prod_{\tau=1}^t p(y_\tau|\mathbf{w}_\tau, \mu, \beta) p(\mathbf{w}_\tau|\mathbf{w}_{\tau-1}, \alpha) \right] p(\theta|\omega), \quad (10)$$

where  $p(y_\tau|\mathbf{w}_\tau, \mu, \beta) = \mathcal{N}(y_\tau|\mathbf{x}_\tau^\top \mathbf{w}_\tau + \mu, \beta^{-1})$  and  $p(\mathbf{w}_\tau|\mathbf{w}_{\tau-1}, \alpha) = \mathcal{N}(\mathbf{w}_\tau|\mathbf{w}_{\tau-1}, \alpha^{-1} \mathbf{I}_I)$ .

### 3 Genuinely Online Variational Inference

An exact implementation of Bayesian LGSSMs is formally intractable [8]. Besides sampling methods [9, 10], VB approximations [11, 12] are popular approximate treatments in this context. Nonetheless, the drawback of such VB procedures is that they all require a full pass through the data at each iteration, rendering them impracticable for streaming data. To remedy this, we develop *Genuinely Online Variational Inference* (GOVI), a novel framework whereby VB may be efficiently deployed in the streaming setting, without the need to revisit past data or have advance knowledge of future data. The rationale behind GOVI is to store the joint BYPASS distribution learned on round  $t-1$  so as to recycle it in the subsequent round. This simple principle is reflected by the following probabilistic recursions:

$$p(y_{1:t-1}, \mathbf{w}_{1:t-1}|\langle \theta \rangle_{1:t-1}) = \prod_{\tau=1}^{t-1} p(y_\tau|\mathbf{w}_\tau, \langle \mu \rangle_\tau, \langle \beta \rangle_\tau) p(\mathbf{w}_\tau|\mathbf{w}_{\tau-1}, \langle \alpha \rangle_\tau), \quad (11)$$

$$p(y_{1:t}, \mathbf{w}_{1:t}|\theta, \langle \theta \rangle_{1:t-1}) = p(y_t|\mathbf{w}_t, \mu, \beta) p(\mathbf{w}_t|\mathbf{w}_{t-1}, \alpha) p(y_{1:t-1}, \mathbf{w}_{1:t-1}|\langle \theta \rangle_{1:t-1}), \quad (12)$$

where  $\langle \theta \rangle_t \equiv \langle \theta \rangle_{q_t(\theta)}$ ,  $\langle \cdot \rangle_{d(x)}$  denotes the expectation w.r.t. the distribution  $d(x)$ , and  $q_t(\cdot)$  is a shorthand for the approximating density  $q(\cdot|y_{1:t}, \langle \theta \rangle_{1:t-1})$ . A crucial implication of this recycling process is that we may discard observations after processing them. As a result, GOVI is both single-pass and computationally efficient, thereby achieving the desiderata of streaming methods [13].

<sup>4</sup>We use a condensed integral notation: all integrals are definite integrals over the entire domain of interest.

<sup>5</sup>A fully Bayesian treatment certainly requires the specification of a *hyperprior*, but is not taken here for space restrictions.

<sup>6</sup> $v_{1:t}$  denotes  $v_1, \dots, v_t$ .

To determine  $q_t(\cdot)$ , one considers the lower bound:

$$\log p(y_{1:t} | \langle \theta \rangle_{1:t-1}, \omega) \geq \langle E_t(\mathbf{w}_{1:t}, \theta) \rangle_{q_t(\mathbf{w}_{1:t}, \theta)} + \langle \log p(\theta | \omega) \rangle_{q_t(\theta)} + H(q_t) \equiv \mathcal{L}, \quad (13)$$

where  $E_t(\mathbf{w}_{1:t}, \theta) \equiv \log p(y_{1:t}, \mathbf{w}_{1:t} | \theta, \langle \theta \rangle_{1:t-1})$  and  $H(d)$  signifies the entropy of  $d(x)$ . The key approximation in VB, commonly called the mean-field approximation (MFA), is  $q_t(\mathbf{w}_{1:t}, \theta) = q_t(\mathbf{w}_{1:t}) \prod_i q_t(\theta_i)$ , from which one may show that, for optimality of  $\mathcal{L}$ ,

$$q_t(\mathbf{w}_{1:t}) \propto q(y_{1:t}, \mathbf{w}_{1:t}) \equiv e^{\langle E_t(\mathbf{w}_{1:t}, \theta) \rangle_{q_t(\theta)}}, \quad q_t(\theta) \propto p(\theta | \omega) e^{\langle E_t(\mathbf{w}_{1:t}, \theta) \rangle_{q_t(\mathbf{w}_{1:t})}}. \quad (14)$$

These coupled equations need to be iterated to convergence. Our main concern is with the update for  $q_t(\mathbf{w}_t)$ , for which this paper makes a departure from treatments previously developed [11, 12]. We will present final results only, and refer the reader to the Supplementary Material for detailed derivations.

### 3.1 Approximate Filtering

From Eqs. (11)-(12), it follows that

$$q(y_{1:t}, \mathbf{w}_{1:t}) = \prod_{\tau=1}^t \underbrace{\mathcal{N}(y_\tau | \mathbf{x}_\tau^\top \mathbf{w}_\tau + \langle \mu \rangle_\tau, \langle \beta \rangle_\tau^{-1})}_{=q(y_\tau | \mathbf{w}_\tau) \approx p(y_\tau | \mathbf{w}_\tau, \mu, \beta)} \times \underbrace{\mathcal{N}(\mathbf{w}_\tau | \mathbf{w}_{\tau-1}, \langle \alpha \rangle_\tau^{-1} \mathbf{I}_I)}_{=q(\mathbf{w}_\tau | \mathbf{w}_{\tau-1}) \approx p(\mathbf{w}_\tau | \mathbf{w}_{\tau-1}, \alpha)}. \quad (15)$$

Clearly, the above represents the joint distribution of the BYPASS model with sequentially updated, averaged parameters. Thus, inference can be performed using the standard Kalman filter (KF) equations [14, 15]. A direct consequence is that the approximate filtering distribution is Gaussian:

$$q_t(\mathbf{w}_t) = \mathcal{N}(\mathbf{w}_t | \boldsymbol{\mu}_t^{\mathbf{w}}, \boldsymbol{\Sigma}_t^{\mathbf{w}}). \quad (16)$$

The moments of this distribution are iteratively updated as described in Algorithm 1.

### 3.2 Mean Variational Parameters

#### Update for $\alpha$

The approximate posterior over the weight precision is a Gamma distribution whose mean can be found from the following fixed-point iteration:

$$\langle \alpha \rangle_t^{\text{new}} = \frac{2a}{2b + \|\boldsymbol{\mu}_t^{\mathbf{w}} - \boldsymbol{\mu}_{t-1}^{\mathbf{w}}\|_2^2 + \text{tr}(\boldsymbol{\Sigma}_t^{\mathbf{w}} - \boldsymbol{\Sigma}_{t-1}^{\mathbf{w}})}. \quad (17)$$

#### Update for $\beta$

The variational posterior of  $\beta$  is a generalised inverse Gaussian distribution defined by

$$q_t(\beta) = \mathcal{GIG}(\beta | -1, 1, \rho_t), \quad \rho_t = (y_t - \mathbf{x}_t^\top \boldsymbol{\mu}_t^{\mathbf{w}} - \langle \mu \rangle_t)^2 + \mathbf{x}_t^\top \boldsymbol{\Sigma}_t^{\mathbf{w}} \mathbf{x}_t + \hat{V}_t^\mu, \quad (18)$$

where  $\hat{V}_t^\mu$  denotes the variance of  $\mu$  under  $q_t$ . The corresponding update equation is therefore

$$\langle \beta \rangle_t^{\text{new}} = \frac{\mathfrak{K}_0(\sqrt{\rho_t})}{\sqrt{\rho_t} \mathfrak{K}_1(\sqrt{\rho_t})}, \quad (19)$$

where  $\mathfrak{K}_\nu(\cdot)$  denotes the modified Bessel function of the second kind, with index  $\nu$ .

#### Update for $\mu$

The approximating density for  $\mu$  is somewhat intractable and non-standard, but is roughly equal to a truncated Gaussian with lower and upper truncation values of  $-\epsilon$  and  $\epsilon$ , respectively, so we set

$$q_t(\mu) = \mathcal{N}_{[-\epsilon, \epsilon]}^{\text{trunc}}(\mu | y_t - \mathbf{x}_t^\top \boldsymbol{\mu}_t^{\mathbf{w}}, \langle \beta \rangle_t^{-1}). \quad (20)$$

From this, we obtain the following fixed-point equation in  $\langle \mu \rangle_t$ :

$$\langle \mu \rangle_t^{\text{new}} = y_t - \mathbf{x}_t^\top \boldsymbol{\mu}_t^{\mathbf{w}} + \frac{\phi(l_t) - \phi(u_t)}{\sqrt{\langle \beta \rangle_t^{\text{old}} [\Phi(u_t) - \Phi(l_t)]}}, \quad (21)$$

where

$$l_t = \sqrt{\langle \beta \rangle_t} [-\epsilon - (y_t - \mathbf{x}_t^\top \boldsymbol{\mu}_t^{\mathbf{w}})], \quad u_t = \sqrt{\langle \beta \rangle_t} [\epsilon - (y_t - \mathbf{x}_t^\top \boldsymbol{\mu}_t^{\mathbf{w}})], \quad (22)$$

while  $\phi(\cdot)$  and  $\Phi(\cdot)$  denote the PDF and CDF of a standard Gaussian, respectively. Similarly,

$$(\hat{V}_t^\mu)^{\text{new}} = \frac{1}{\langle \beta \rangle_t^{\text{old}}} \left[ 1 + \frac{l_t \phi(l_t) - u_t \phi(u_t)}{\Phi(u_t) - \Phi(l_t)} - \left( \frac{\phi(l_t) - \phi(u_t)}{\Phi(u_t) - \Phi(l_t)} \right)^2 \right]. \quad (23)$$

### 3.3 Relation to Streaming Variational Bayes

In this section, we argue that GOVI falls under a broader family of online VB algorithms known as Streaming Variational Bayes (SVB) [3]. Note that Bayes' rule can be written in a streaming form:

$$p(\Theta|y_{1:t}) \propto p(y_t|\Theta) p(\Theta|y_{1:t-1}), \quad (24)$$

where  $\Theta$  represents a set of stochastic parameters. SVB suggests that, when the above is infeasible to compute, one should adopt an approximation algorithm  $\mathcal{A}$  such that

$$p(\Theta|y_{1:t}) \approx q_t(\Theta) = \mathcal{A}(y_t, q_{t-1}(\Theta)), \quad (25)$$

with  $q_0(\Theta) = p(\Theta)$ . When  $\mathcal{A}$  generates the posterior from Bayes' theorem, this calculation is exact. In the setting of BYPASS,  $\Theta = \{\mathbf{w}_t, \boldsymbol{\theta}\}$  and, by MFA, we obtain two separate approximation algorithms, namely

$$q_t(\mathbf{w}_t) = \mathcal{A}_{\mathbf{w}}(y_t, q_{t-1}(\mathbf{w}_t)) \quad \text{and} \quad q_t(\boldsymbol{\theta}) = \mathcal{A}_{\boldsymbol{\theta}}(y_t, p(\boldsymbol{\theta}|\boldsymbol{\omega})), \quad (26)$$

the latter having to ineluctably rely on a time-invariant prior over  $\boldsymbol{\theta}$ , as the BYPASS framework does not specify any dynamics in that regard. More precisely, we have

$$q_t(\mathbf{w}_t) \propto \mathcal{N}(y_t | \mathbf{x}_t^\top \mathbf{w}_t + \langle \mu \rangle_t, \langle \beta \rangle_t^{-1}) \underbrace{\int \mathcal{N}(\mathbf{w}_t | \mathbf{w}_{t-1}, \langle \alpha \rangle_t^{-1} \mathbf{I}_I) q_{t-1}(\mathbf{w}_{t-1}) d\mathbf{w}_{t-1}}_{=q_{t-1}(\mathbf{w}_t)} \quad (27)$$

$$q_t(\boldsymbol{\theta}) \propto q_t(y_t | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \boldsymbol{\omega}), \quad q_t(y_t | \boldsymbol{\theta}) = \exp \left\{ \langle \log p(y_t | \mathbf{w}_t, \mu, \beta) p(\mathbf{w}_t | \mathbf{w}_{t-1}, \alpha) \rangle_{q_t(\mathbf{w}_{1:t})} \right\}. \quad (28)$$

Interestingly, from Eq. (27), we are able to recover the KF equations evaluated at the mean variational parameters. The aforementioned digression from treatments previously presented thus emanates from the fact that we make the first application of SVB to the Bayesian LGSSM setting.

## 4 Learning the hyperparameters: Adaptive BYPASS

As far as variational inference in Bayesian LGSSMs is concerned, the optimal hyperparameter values are typically obtained by optimising the variational lower bound  $\mathcal{L}$  w.r.t. to  $\boldsymbol{\omega}$  [11, 12]. However, this would not be computationally viable in a streaming environment. Since we are not treating  $\boldsymbol{\omega}$  as a random vector, we may readily apply the PA regression framework from Section 2.1 to automatically tune  $\boldsymbol{\omega}$  in an online manner. To mimic the ML-II ('evidence') framework, we use the negative log likelihood of the BYPASS model as the underlying loss function. This gives rise to the following optimisation problem:

$$\hat{\boldsymbol{\omega}}_t = \arg \min_{\boldsymbol{\omega} > \mathbf{0}_{M \times 1}} \left\{ \frac{1}{2} \|\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}_{t-1}\|_2^2 + C_{\boldsymbol{\omega}} \frac{\beta}{2} (y_t - \mathbf{x}_t^\top \boldsymbol{\mu}_{t-1}^{\mathbf{w}} - \mu)^2 \right\}, \quad (29)$$

where  $M \equiv \dim(\boldsymbol{\omega})$ . We remark that, by construction, this problem corresponds to sequential maximum likelihood at the hyperparameter level. Its objective function depends on  $\boldsymbol{\omega}$ , insofar as the latter is employed to determine the weight estimates  $\boldsymbol{\mu}_{t-1}^{\mathbf{w}}$ . To convert this problem into a more 'conventional' one, we replace the strict-positivity constraints  $\boldsymbol{\omega} > \mathbf{0}_{M \times 1}$  by  $\boldsymbol{\omega} \geq \boldsymbol{\omega}_{\min}$ , where  $\boldsymbol{\omega}_{\min} \approx \mathbf{0}_{M \times 1}$  represents a lower bound on  $\boldsymbol{\omega}$ . We consequently get (see Supplementary Material)

$$\hat{\boldsymbol{\omega}}_t = \max \left\{ \hat{\boldsymbol{\omega}}_{t-1} + C_{\boldsymbol{\omega}} \langle \beta \rangle_{t-1} \mathbf{x}_t^\top \boldsymbol{\psi}_{t-1} (y_t - \mathbf{x}_t^\top \boldsymbol{\mu}_{t-1}^{\mathbf{w}} - \langle \mu \rangle_{t-1}) \mathbf{1}_{M \times 1}, \boldsymbol{\omega}_{\min} \right\}, \quad (30)$$

where the max operator is taken element-wise,  $\mathbf{1}_{D \times 1}$  denotes a  $D$ -dimensional vector of ones and, for each  $\omega \in \boldsymbol{\omega}$ ,  $\psi_t$  denotes the gradient of  $\boldsymbol{\mu}_t^{\mathbf{w}}$  w.r.t.  $\omega$  evaluated at  $\omega = \hat{\omega}_t$ . As demonstrated in the Supplementary Material, this gradient is updated in an iterative fashion, based on its previous value and  $\mathbf{S}_t$ , the gradient of  $\boldsymbol{\Sigma}_t^{\mathbf{w}}$  w.r.t.  $\omega$  evaluated at  $\hat{\omega}_t$ . We dubbed the ensuing algorithm *adaptive BYPASS* (ADA-BYPASS). The implementation details of the latter and of its non-adaptive counterpart are outlined in Algorithms 2 and 1, respectively.

---

**Algorithm 1** BYPASS

---

- 1: **Input:** Hyperparameters  $\boldsymbol{\omega}$ , initial mean variational parameters  $\langle \boldsymbol{\theta} \rangle_0$ .
  - 2: Set  $\boldsymbol{\mu}_0^{\mathbf{w}} = \mathbf{0}_{I \times 1}$  and  $\boldsymbol{\Sigma}_0^{\mathbf{w}} = \mathbf{0}_{I \times I}$ .
  - 3: **for**  $t = 1, 2, \dots$  **do**
  - 4:   Obtain new inputs  $\mathbf{x}_t$ .
  - 5:   Compute the predictive mean and variance of the output:
$$\hat{m}_t = \mathbf{x}_t^\top \boldsymbol{\mu}_{t-1}^{\mathbf{w}} + \langle \mu \rangle_{t-1}, \quad \hat{V}_t = \mathbf{x}_t^\top \mathbf{P}_{t-1}^{\mathbf{w}} \mathbf{x}_t + \langle \beta \rangle_{t-1}^{-1}.$$
  - 6:   Derive the new mean variational parameters  $\langle \boldsymbol{\theta} \rangle_t$  by repeating the fixed-point iterations (17), (19), (21) and (23) until convergence.
  - 7:   Evaluate the *predictive weight covariance* and the *Kalman gain*:
$$\mathbf{P}_{t-1}^{\mathbf{w}} = \boldsymbol{\Sigma}_{t-1}^{\mathbf{w}} + \langle \alpha \rangle_t^{-1} \mathbf{I}_I, \quad \mathbf{g}_t = (\mathbf{x}_t^\top \mathbf{P}_{t-1}^{\mathbf{w}} \mathbf{x}_t + \langle \beta \rangle_t^{-1})^{-1} \mathbf{P}_{t-1}^{\mathbf{w}} \mathbf{x}_t.$$
  - 8:   Update the mean and covariance of the approximate filtering distribution  $q_t(\mathbf{w}_t)$ :
$$\boldsymbol{\mu}_t^{\mathbf{w}} = \boldsymbol{\mu}_{t-1}^{\mathbf{w}} + \mathbf{g}_t (y_t - \mathbf{x}_t^\top \boldsymbol{\mu}_{t-1}^{\mathbf{w}} - \langle \mu \rangle_t), \quad \boldsymbol{\Sigma}_t^{\mathbf{w}} = (\mathbf{I}_I - \mathbf{g}_t \mathbf{x}_t^\top) \mathbf{P}_{t-1}^{\mathbf{w}}.$$
  - 9: **end for**
- 

---

**Algorithm 2** ADA-BYPASS: BYPASS with hyperparameter adaptation via PA regression.

---

- 1: **Input:** Initial hyperparameters  $\hat{\omega}_0$ , lower hyperparameter bounds  $\boldsymbol{\omega}_{\min}$ , initial mean variational parameters  $\langle \boldsymbol{\theta} \rangle_0$ , initial variational variance  $\hat{V}_0^\mu$ , aggressiveness parameter  $C_\omega > 0$ .
  - 2: Same as Step 2 in Algorithm 1.
  - 3: Initialise the gradients w.r.t.  $\omega \in \boldsymbol{\omega}$ :  $\boldsymbol{\psi}_0 = \mathbf{0}_{I \times 1}$ ,  $\mathbf{S}_0 = \mathbf{I}_I$ .
  - 4: **for**  $t = 1, 2, \dots$  **do**
  - 5:   Same as Steps 4-5 in Algorithm 1.
  - 6:   Update the hyperparameters according to Eq. (30).
  - 7:   Same as Steps 6-8 in Algorithm 1.
  - 8:   Update the gradients:
$$\mathbf{S}_t = (\mathbf{I}_I - \mathbf{g}_t \mathbf{x}_t^\top) \mathbf{S}_{t-1} (\mathbf{I}_I - \mathbf{x}_t \mathbf{g}_t^\top),$$

$$\boldsymbol{\psi}_t = (\mathbf{I}_I - \mathbf{g}_t \mathbf{x}_t^\top) \boldsymbol{\psi}_{t-1} + \langle \beta \rangle_t \mathbf{S}_t \mathbf{x}_t (y_t - \mathbf{x}_t^\top \boldsymbol{\mu}_{t-1}^{\mathbf{w}} - \langle \mu \rangle_t).$$
  - 9: **end for**
- 

## 5 Applications

### 5.1 Practicalities

Based on the sensitivity analysis in [1], we set the aggressiveness parameter  $C_\omega$  equal to  $10^{-3}$ . The model parameters are initialised at their prior means, except for the output precision  $\beta$ , whose prior mean is undefined. A similar principle is applied to the variational variance of  $\mu$ . As a result of this, we obtain:  $\langle \alpha \rangle_0 = a/b$ ,  $\langle \mu \rangle_0 = 0$  and  $\hat{V}_0^\mu = \text{Var}[\overline{\mathcal{U}}(\mu) - \epsilon, \epsilon] = \epsilon^2(1 + \epsilon/3)/(1 + \epsilon)$ . As for  $\beta$ , we approximate its prior mean as follows:  $\langle \beta \rangle_0 \approx 0.5/10^{-3} = 500$ .

Next, we choose initial values for the hyperparameters. In order to initially emulate the frequentist PA regression framework (Section 2.1) while simultaneously making  $p(\alpha|a, b)$  ‘uninformative’ (i.e. broad), we set  $a = C_\omega^{-1}$  and  $b = 1$ . As for the insensitivity hyperparameter, we use  $\epsilon = 1.25$ , this

value being the mean of a symmetric Beta distribution of the second kind<sup>7</sup> with shape parameter  $s = 5$ , the choice of which was motivated by [16]. Finally, we selected  $\omega_{\min} = 10^{-8}$ ,  $\forall \omega_{\min} \in \omega_{\min}$ .

## 5.2 Model specification and benchmark

In the following experiments, unless otherwise stated, we used an autoregressive measurement equation of order 1 (AR(1)):  $y_t = w_{t,0} + w_{t,1}y_{t-1} + \eta_t$ , where  $w_{t,0}$  is a bias parameter. While this is perhaps not the best specification, feature selection goes beyond the scope of the present study. It is worthwhile noting, however, that there is no theoretical or practical obstacle that would prevent us from considering more complex predictors. This would be expected to further improve the model's performance.

We make comparisons with a standard LGSSM in which a MAP recursion is used to govern the adaptation of the model parameters, by sequentially using the maximum-likelihood formulation first proposed by [17]. To ensure full comparability of results, we also endow this model with an AR(1) hypothesis, and refer to it as *sequential Kalman filter* (SKF) in the applications below.

In both models, one-step ahead forecasts are successively iterated to provide multi-step forecasts of arbitrary length, as needed. Missing values, if they occur, are accommodated for using the scheme advocated by [5], in which they are replaced by their expectations under the corresponding model.

## 5.3 Nile data

We first consider a canonical changepoint data set, the minimum water levels of the Nile river during the period AD 622-1284 [18]. Several authors have found evidence supporting a changepoint for these data around AD 720-722 [18, 19, 20]. The conjectured reason for this changepoint is the construction in AD 715 of a new device (a 'nilometer') on the island of Roda, which affected the nature and accuracy of the measurements.

We performed one-year lookahead prediction on this data set. The results can be seen in Fig. 1. We note the superior performance of ADA-BYPASS compared with the SKF.

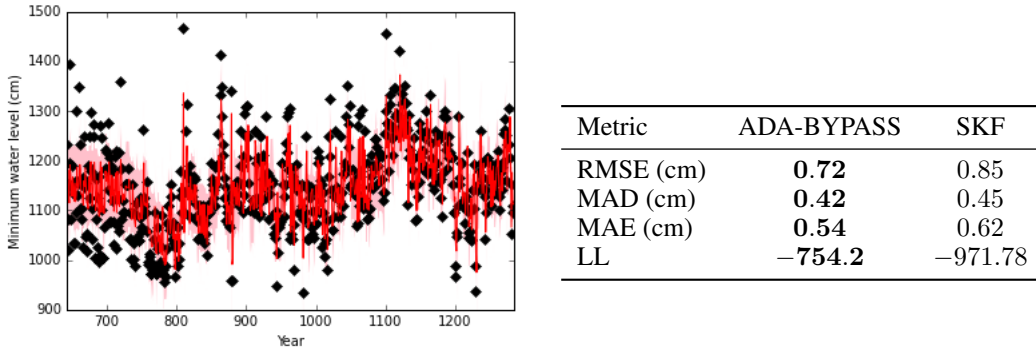


Figure 1: Online one-year ahead predictions for the Nile's minimum water levels. Left panel: observed levels (black diamonds), predicted levels (red line) and  $\pm 1$  standard deviation error bars (pink area). Right panel: predictive performances; error metrics shown are root mean squared error (RMSE), mean absolute deviation (MAD), mean absolute error (MAE) and predictive log likelihood (LL).

## 5.4 Wind speed data

To demonstrate the superior performance of ADA-BYPASS on a large data set, we next present the series of anemometer wind speed measurements (in m/s) from a Danish wind turbine. The data were sampled at 10 minute intervals for just over nine months, resulting in a total of 40,174 measurements. The 10 minute lookahead predictive performance achieved by each method is reported in Table 1.

<sup>7</sup>We show in the Supplementary Material that the form of  $p(\eta_t|\epsilon)$  induces this prior for  $\epsilon$ .

Table 1: Predictive performance of ADA-BYPASS vs SKF on the wind speed data set.

Metric	ADA-BYPASS	SKF
RMSE (m/s)	<b>0.6</b>	0.64
MAD (m/s)	<b>0.3</b>	0.31
MAE (m/s)	<b>0.42</b>	0.44
LL	<b>−24,971.75</b>	−30,140.04

## 5.5 Statistical Arbitrage

LGSSMs, and variants thereof, have seen a widespread use in statistical arbitrage strategies, notably in pairs trading [21, 22, 23]. In this area, they serve as a dynamic model for the price spread between two assets. In our application, we seek to find the *hedge ratio*<sup>8</sup> and the predictive standard deviation of the spread. The observable variable is thus one of the price series  $y$ , and the hidden variable is the hedge ratio  $w$ . We assume that both variables obey the ADA-BYPASS dynamics, i.e.

$$y_t = w_t x_t + \eta_t, \quad \eta_t \sim \mathcal{N}(\eta_t | \mu, \beta^{-1}), \quad w_t = w_{t-1} + \zeta_t, \quad \zeta_t \sim \mathcal{N}(\zeta_t | 0, \alpha^{-1}), \quad (31)$$

where  $x$  is the price series of the other asset. Typically,  $\alpha$ ,  $\beta$  and  $\mu$  are manually selected in hindsight [21]. However, this practice is highly prone to the so-called *data-snooping bias*: these parameters can be tweaked so as to optimise the backtesting performance of the strategy. The ADA-BYPASS algorithm automatically tunes its underlying parameters, so it does not suffer from this caveat.

We tested ADA-BYPASS on a pair of exchange-traded funds (ETFs) consisting of the SPDR gold trust GLD and the gold-miners ETF GDX. This ETF pairing is a favourite in the financial industry, because the value of gold-mining companies is very much based on the value of gold. We downloaded the corresponding, daily adjusted closing prices from Yahoo! Finance, between 22/05/2006 and 22/04/2015.

Rather than maximising profits, most investors attempt to maximise risk-adjusted returns, as advocated by modern portfolio theory. The Sharpe ratio is the most widely used measure of risk-adjusted returns [24]. Besides the Sharpe ratio, the maximum drawdown and maximum drawdown duration are two other popular metrics to evaluate trading strategies. From Table 2, we can clearly discern that ADA-BYPASS beats SKF by a significant margin in terms of the aforementioned performance metrics.

Table 2: Performance of the GDX-GLD pairs trade under ADA-BYPASS and SKF.

Metric	ADA-BYPASS	SKF
Sharpe ratio	<b>1.12</b>	0.7
Maximum drawdown (%)	<b>14.61</b>	73.05
Maximum drawdown duration (trading days)	<b>375</b>	567

## 6 Concluding remarks

We introduced the first online Bayesian PA regression model within the state-space setting, along with a novel, online variational inference algorithm. This model is ideal for the probabilistic prediction of non-stationary and/or very large time series, in particular massive, time-varying data streams. Results on three real-world data sets show significant improvements in predictive performance over a more standard LGSSM.

<sup>8</sup>The hedge ratio of a particular asset is the number of units of that asset we should buy or sell in a portfolio. If the asset is a stock, then the number of units corresponds to the number of shares. A negative hedge ratio indicates we should sell that asset.



## References

- [1] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online Passive-Aggressive Algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.
- [2] F. Hutter, H. Hoos, and K. Leyton-Brown. An Efficient Approach for Assessing Hyperparameter Importance. In T. Jebara and E. P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, volume 32, pages 754–762. JMLR Workshop and Conference Proceedings, 2014.
- [3] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. Streaming Variational Bayes. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 1727–1735. Curran Associates, Inc., 2013.
- [4] M. S. Grewal and A. P. Andrews. *Kalman Filtering: Theory and Practice Using MATLAB*. John Wiley and Sons, Inc., 4th edition, 2015.
- [5] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer-Verlag New York, 3rd edition, 2011.
- [6] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.
- [7] M. Pontil, S. Mukherjee, and F. Girosi. On the Noise Model of Support Vector Machines Regression. In H. Arimura, S. Jain, and A. Sharma, editors, *Algorithmic Learning Theory*, volume 1968 of *Lecture Notes in Computer Science*, pages 316–324. Springer Berlin Heidelberg, 2000.
- [8] M. Davy and S. J. Godsill. Bayesian Harmonic Models for Musical Signal Analysis. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 105–124. Oxford University Press, 2003.
- [9] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer-Verlag New York, 2005.
- [10] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer-Verlag New York, 2006.
- [11] D. Barber and S. Chiappa. Unified Inference for Variational Bayesian Linear Gaussian State-Space Models. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, pages 81–88. MIT Press, 2007.
- [12] S. Chiappa and D. Barber. Bayesian Factorial Linear Gaussian State-Space Models for Biosignal Decomposition. *IEEE Signal Processing Letters*, 14(4):267–270, 2007.
- [13] P. Domingos and G. Hulten. A General Framework for Mining Massive Data Streams. *Journal of Computational and Graphical Statistics*, 12(4):945–949, 2003.
- [14] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the American Society for Mechanical Engineering, Series D, Journal of Basic Engineering*, 82:35–45, 1960.
- [15] P. Zarchan and H. Musoff. *Fundamentals of Kalman Filtering: A Practical Approach*. American Institute of Aeronautics and Astronautics (AIAA), 3rd edition, 2009.
- [16] S. J. Roberts and W. D. Penny. Variational Bayes for Generalized Autoregressive Models. *IEEE Transactions on Signal Processing*, 50(9):2245–2257, 2002.
- [17] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.
- [18] B. Whitcher, S. D. Byers, P. Guttorp, and D. B. Percival. Testing for homogeneity of variance in time series: Long memory, wavelets, and the Nile River. *Water Resources Research*, 38(5):12–1–12–16, 2002.
- [19] R. Garnett, M. A. Osborne, S. Reece, A. Rogers, and S. J. Roberts. Sequential Bayesian Prediction in the Presence of Changepoints and Faults. *The Computer Journal*, 53(9):1430–1446, 2010.
- [20] B. K. Ray and R. S. Tsay. Bayesian methods for change-point detection in long-range dependent processes. *Journal of Time Series Analysis*, 23(6):687–705, 2002.
- [21] E. P. Chan. *Algorithmic Trading: Winning Strategies and their Rationale*. Wiley Trading Series. John Wiley and Sons, Inc., 2013.
- [22] K. Triantafyllopoulos and G. Montana. Dynamic modeling of mean-reverting spreads for statistical arbitrage. *Computational Management Science*, 8(1–2):23–49, 2011.
- [23] G. Montana, K. Triantafyllopoulos, and T. Tsagaris. Flexible least squares for temporal data mining and statistical arbitrage. *Expert Systems with Applications*, 36(2):2819–2830, 2009.
- [24] W. F. Sharpe. Mutual Fund Performance. *The Journal of Business*, 39(1):119–138, 1966.